COMPARISON OF FOUR- AND FIVE-OPTION MULTIPLE-CHOICE QUESTIONS IN NURSING ENTRANCE TESTS

M. Panczyk, H. Rebandel, J. Gotlib

Division of Teaching and Outcomes of Education, Faculty of Health Sciences, Medical University of Warsaw (POLAND)

Abstract

Introduction:

Most multiple-choice tests comprise questions with four options per item. However, a number of academic teachers believe that a larger number of options per question shall increase the scope of variability of test results. An increase in discrimination capability is particularly important with reference to selective examinations. In 2011 and 2013, Nursing Entrance Test at Medical University of Warsaw (MUW) comprised 5-option items tests, which was an exception from 4-option items tests used in 2009-2010 and 2012.

Aim of study:

Assessment of the impact of change of the number of options in multiple-choice questions (MCQs) on the quality of Nursing Entrance Exams for an MA programme at MUW between 2009-2013.

Materials and Methods:

A total of 250 multiple-choice exam questions, including 150 four-option items (2009-2010 and 2012) and 100 five-option items (2011 and 2013). In order to compare the quality of particular exams, the level of easiness, substitute differentiation power, and Pearson's linear correlation coefficient were established for each pool of questions. The comparison comprised the scope of variability of the results (coefficient of variation, scope of results, and quartile range) as well as the average easiness and capacity of differentiating particular questions in consecutive versions of the exam. A one-way analysis of ANOVA variance and *post-hoc* RIR Tukey honestly significant difference test were used.

Results:

In 2011 and 2013, when the 5-option items tests were introduced, the difficulty of the exam expressed as the mean score amounted to 24.3 and 25.7 points, respectively. These values are comparable to the results achieved in 2010 (25.6), but they are clearly different from those obtained in 2009 (30.2) and 2012 (31.5). Similar differences were observed in comparison of coefficients of variation that were similar in 2010, 2011, and 2013 (17.1, 17.9 and 18.4%, respectively) and significantly different from those obtained in 2009 and 2012 (14.3 and 14.1%, respectively). Moreover, a greater symmetry (skewness ≈ 0) in frequency distribution of test scores was observed in the case of 5-option items tests compared to 4-option items tests. The reliability of the exam was variable, Cronbach's α coefficient ranged between 0.429 and 0.559. No statistically significant differences were found in discrimination capability of the exams performed in the form of 4- or 5-option items tests (ANOVA test, $P > 0.05$). It was also demonstrated that the 2011 exam (5 options) was significantly more difficult than that of 2012 (4 options) (ANOVA test ($P = 0.0025$) and *post-hoc* RIR Tukey honestly significant difference test ($P < 0.01$)).

Conclusions:

The introduction of an additional option item to the test questions did not significantly improve the qualitative parameters of the Nursing Entrance Exams at MUW. Significant increase in selective capacity of the exam and reliability of assessment was not observed. It is recommended to use 4-option items tests and to develop a good test content outline for future editions of the exam.

Keywords: educational measurement, nursing admission examination, nursing entrance test, number of choices per item, item discrimination.

1 INTRODUCTION

Using multiple-choice questions (MCQs) in medical and nursing education is one of the most popular forms of checking knowledge and skills of the examinees. Despite certain critical opinions concerning the usefulness of the MCQs tests in the evaluation of clinical abilities [1] and the candidates' predispositions to become students [2], this method still presents many advantages as opposed to other tools used in didactics, such as short answer or essay-style questions. Well constructed MCQs allow not only to assess the ability to simply recall the memorised facts but also to measure the abilities of practical application of this knowledge so as to solve certain clinical problems [3]. Thus, using MCQs in competence tests may serve well in thorough selection, differentiating those who achieve higher and more clear scores in a given domain of knowledge and skills as opposed to the rest who take the same exam [4]. Additionally, MCQs are thought to be more objective and allow the teacher to check a broader spectre of competences than it is the case with other written forms of evaluation. Also, using MCQs allows an efficient and quick assessment of a large group of candidates and the obtained results of exams may also undergo a reliability and item analysis as well as standardization [5, 6].

The available test banks and course books for nursing education include mostly tests based on MCQs prepared in form of four-option (3 distractors and 1 correct answer) [7, 8]. Even though the experts dealing with psychometric analyses have been underlining for many years that the three-option one is the most optimal number of items in examination papers [9, 10], still majority of teachers use items with four or five options [11, 12]. MCQs built on a lower number of options are much less popular in the academic environment despite presenting some significant advantages. A lower number of MCQs allows to select alternative answers better and creating such a task takes much less time. Detailed analyses of functioning of individual distractors in question that have four or more options show that the examinees really rarely choose answers from the whole list of options, most of the time they focus on just two or three available distractors [11, 12]. Moreover, creating a test set on the basis of three-option items gives a possibility to enlarge a measuring scale (which increases the overall reliability of the measurement) but without an excessive increase in time needed to prepare such an exam [11].

As part of Nursing Entrance Test for the second level of studies (MA studies), candidates are tested on their knowledge and skills in four areas: *basic science*, *health sciences*, *primary healthcare,* and *clinical nursing*. Such a broad range of subjects and the differentiated range of the evaluated competences required applying good selective tools that would objectively and quickly assess a large group of candidates for university. Following Tarrant & Ware (2010), in case of MCQs exams, we are faced with high efficiency and ability to assess various results of education and thus this form of examining is still one of the best methods of assessment in nursing [13]. Between the years 2009 – 2013, at Medical University of Warsaw (MUW) there were five editions of an entrance exam in form of a test, which was the main discrimination criterion. As in the case of other methods used in educational measurement, it is indispensable that the quality of Nursing Entrance Test be permanently improved, so as to fulfil all requirements for such a tool. Furthermore, if the time required to develop multiple-choice tests can be reduced without reducing the reliability and validity of the assessment, then in case of confirmed data from the retrospective analyses, the department team preparing the entrance exam will be able to recommend certain changes in the structure of the test. Carrying out a flexible recruitment strategy that would be in agreement with the concept of e*vidence-based admissions criteria* is necessary so as to obtain an effective way of selecting the best candidates for whom it will be possible to forecast with high probability that, on the one hand, they will achieve good scores while studying and on the other hand, they will achieve a professional success on higher positions of nursing management [14].

The aim of the results presented here was to assess the influence of a change in a number of options if MCQs on the quality of Nursing Entrance Tests during the course of studies at the second level of studies at MUW between the years 2009-2013.

2 MATERIALS AND METHODS

The presented observational study is a 5-year long retrospective analysis. The study used the admission data of candidates for a Master's degree programme in Nursing in 2009-2013 (n = 2257, median of age = 23 years). Each MCQs test comprised 50 questions in the "best answer from a list of possible answers" format. Overall, there were 250 exam questions that were analysed, 150 of which had a form of four options (the years 2009-2010 and 2012), and 100 questions had five options of answers (the years 2011 and 2013).

Raw data were preprocessed using TESTY version 7 (Testy komputerowe, Copyright © 1994-2014 by Sławomir Zalewski) and exported to Statistica version 10 (StatSoft, Inc.) for further analysis. All programs were used in compliance with the MUW license.

Normal distribution parameters of particular exam results were assessed using the Shapiro-Wilk test and data were screened for outliers using Grubbs test. Cronbach's α coefficient was estimated to determine the reliability level of the test [15]. In order to compare the quality of individual exams for every set of questions, a level of item difficulty was established as well as item discrimination, and r-Pearson linear coefficient. The range of changes of the obtained results was compared (variances, coefficient of variation, range of results and quartile range), but also an average simplicity and the ability to differentiate individual items in subsequent editions of the exam. ANOVA test was used together with a *post-hoc* RIR Tukey's sensible significant difference test.

For all analyses, the *a priori* level of significance was 0.05.

3 RESULTS

While analysing the results of individual editions of the MCQs exam considering the character of dispersion of this variable, a slight askew was noted and in some cases a set of result differed from the usual set (kurtosis ≠ 0 and Shapiro-Wilk test, $P< 0.05$). At the same time, no presence of outlying results was observed (Grubbs test, $P> 0.05$). Individual editions of MCQs exams differed in the range of result changeability which was reflected in different values of coefficients, result hiatus and the range of score obtained by the candidates in individual years. In 2011 and 2013, when the five-option questions were introduced, the difficulty of the exam measured by the average of obtained points, oscillated around 24.3 and 25.7 points respectively. This is a value comparable with the results from the year 2010 (25.6), although it is significantly different to those from the years 2009 and 2012 (30.2 and 31.5 respectively). Similar differences were noted while comparing the results of a coefficient of variation that were similar in the editions of 2010, 2011 and 2013 (17.1, 17.9 and 18.8% respectively), and they differ significantly from those from the years 2009 and 2012 (14.3 and 14.1% respectively).

Moreover, a greater symmetry was observed (skewness ≈ 0) in the frequency of dispersion of the results in case of exams with five-option items as opposed to those with four-option ones. Narrow confidence interval (CI) for the average and the standard deviation are the evidence of high precision of the assessed parameter ranges for the studied population. All of the presented results were summarised in Table 1.

Table 1. Characteristics of particular editions of test exam.

	Four-option items			**Five-option items**		***P*-value**
	2009	**2010**	**2012**	**2011**	**2013**	
Mean (95% CI)	30.2 (29.9-30.6)	25.6 (25.3-26.0)	31.5 (31.0-31.9)	24.3 (23.9-24.7)	25.7 (25.1-26.2)	----
SD (95% CI)	4.33 (4.08-4.60)	4.59 (4.34-4.87)	4.44 (4.13-4.80)	4.16 (3.92-4.44)	4.71 (4.36-5.13)	0.03*
Median	30.0	25.0	31.0	25.0	26.0	0.0001**
$Q_1 – Q_3$	27.0-33.0	23.0-29.0	29.0-34.0	21.0-27.0	22.0-29.0	----
Range of scores	16.0-42.0	11.0-40.0	15.0-45.0	13.0-37.0	12.0-39.0	----
CV	14.3%	17.9%	14.1%	17.1%	18.4%	----
Skewness	0.046	0.109	-0.112	-0,050	0.064	----
Kurtosis	0.178	-0.005	0.327	0.022	-0.274	----
Cronbach's α	0.558	0.546	0.446	0.429	0.559	----
SEM	2.88	3.09	3.30	3.14	3.13	----

* Levene test for equality of variances

** Kruskal-Wallis non-parametric ANOVA

CI – confidence interval; SEM – standard error of measurement; Q_1 and Q_3 – first and third quartile; SD - standard deviation; CV - coefficient of variation

Analysis of individual editions of the exam shows the insufficient reliability of the set of test questions. In subsequent years, Cronbach's α coefficient was between 0.429 and 0.559. Standard error of measurement oscillated around 3 points in each analysed year (SEM ranging between 2.88 and 3.30). Detailed results of reliability analysis were presented in Table 1.

Comparative analysis of weighted average of values of the discriminating power and r-Pearson's coefficients of linear correlation for the set of test questions in subsequent years showed no differences significant statistically for these quality parameters (Levene test for equality of variances, $P > 0.05$; ANOVA test, $P > 0.05$). However, it appeared that the exam in 2011 (five-option items) was significantly more difficult than the exam in 2012 (four-option items) (one-way analysis of variance ANOVA (Eta-squared = 0.065,$P = 0.0025$) and *post-hoc* RIR Tukey honestly significant difference test ($P < 0.01$)). A detailed summary of comparative analysis for quality parameters for individual editions of the exam were presented in Table 2 and Fig. 1.

Table 2. Quality parameters for particular exam editions.

	Four-option items			**Five-option items**		***P*-value***
	2009	**2010**	**2012**	**2011**	**2013**	
Item difficulty						
Mean (95% CI)	0.605 (0.528-0.682)	0.514 (0.450-0.578)	0.630 (0.598-0.661)	0.486 (0.422-0.550)	0.513 (0.449-0.577)	0.0025
SD	0.2717	0.2262	0.1094	0.2243	0.2252	
Item discrimination						
Mean (95% CI)	0.136 (0.114-0.157)	0.149 (0.126-0.172)	0.140 (0.120-0.160)	0.134 (0.112-0.155)	0.155 (0.127-0.183)	NS
SD	0.0752	0.0820	0.0703	0.0765	0.0983	
Pearson product-moment correlation coefficient						
Mean (95% CI)	0.207 (0.185-0.228)	0.203 (0.176-0.230)	0.190 (0.169-0.211)	0.184 (0.161-0.207)	0.207 (0.176-0.238)	NS
SD	0.0761	0.0954	0.0742	0.0801	0.1100	

* one-way analysis of variance ANOVA

CI – confidence interval; SD - standard deviation

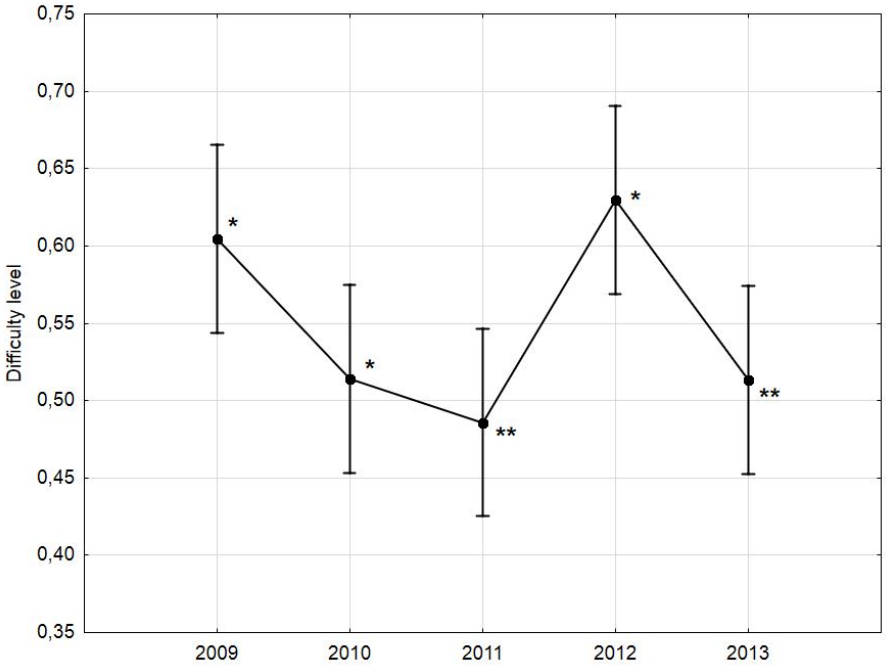

Figure 1. Analysis of variance for difficulty level of tasks from set of exam questions in the subsequent years(*P* = 0.0025; vertical columns show 95% CI for the mean; [*] four-option items, [**] five-option items).

4 DISCUSSION

Findings obtained during the course of analyses remain in agreement with conclusions from similar research that are currently available in literature. Generally, increasing the number of options in MCQs does not have a positive impact on the quality parameters of exam tests. Introducing an additional distractor to test questions did not have a significant influence on the exam discrimination ability despite the increase in difficulty in case of an exam comprising of five-option items. The increase in the level of difficulty resulted in slight narrowing in the range of the obtained results and lowering asymmetry of frequency distribution. A similar result could also be obtained applying a set of four-option items. As results of analysis concerning the results of the exam of 2010 (four-option items), the average median score range and the coefficient of variation was comparable with those obtained for the years 2011 and 2013 (five-option items). A change in the number of options did not improve the quality parameters of the questions, such as differentiating replacement power and r-Pearson's linear correlation coefficient. Both parameters are a measure of the ability of the questions used in the construction of a given exam. A change from four to five options was to increase the differentiating power of the Nursing Entrance Test. However, a detailed analysis of this change's influence on the quality of the tool points out that if in the first year (2011) when a four-option format was used the exam was more difficult, then in 2013 the difficulty level of the exam was comparable with the past editions in which four-option items were included.

An appropriate selection of exam questions considering their level of difficulty is a crucial aspect to not be neglected while constructing a test. A high number of easy questions and approbative ones, i.e. such questions that require of the examinee only confirmation of a given piece of information (e.g., the use of ''all of the above'' (AOTA) and ''none of the above'' (NOTA) as options) are less demanding intellectually and thus contribute to achieving higher average score by candidates who take the same exam [16]. Distribution of result variables will then have the shape of a negative skew and the range of discrimination possibilities will be much narrowed. That is why it is so important to select such exam questions that would show a greater differentiating power. Such questions can distinguish the candidates who achieve results significantly different for a given measured value, which allows a good selection. Such questions are frequently referred to in literature as "ripping items" [17] as they cause flattening if the results distribution (kurtosis < 0), which makes discrimination and determination of the passed / failed cut-off point much easier.

Rogers and Harley (1999) [18] showed the need to carry out detailed analyses concerning lowering the influence of the number of distractors on the MCQs' psychometric parameters. As can be seen from the examples of analyses carried out for medical tests [19, 20] and nursing ones [13], reducing the number of options to three does not influence the quality of test questions significantly. Through reliability and item analysis as well as elimination of the non-functioning distractors from MCQs a number of positions in four- and five-option items can be easily reduced [13, 19, 20]. As was proven by Tarrant and Ware (2010) [13] the differences in item difficulty and discrimination between four-option items and the same items rewritten as three-option items were small and statistically non-significant.

Meta-analysis prepared by Rodriguez (2005) [10], a rather lengthy work, showed that in case of well-prepared MCQs, three-option items are quite enough. The author points out that if with three-option questions the level of difficulty is lowered, however, the differentiating power and reliability increase at the same time [10]. Also the reviews of research published by Vyas and Supe (2008) [9] and Haladyn et.al (2002) [21] confirm the above conclusions. As the authors state, three-option questions are of similar efficiency to four-option ones. However, their major advantage is their easier construction as the creators of questions can suggest distractors of good quality that will differentiate the examines better. Moreover, there is also lower likelihood of the used options to be presenting such serious construction flaws that the examinees will not even consider them as equivalent choices [17, 22, 23]. Valuating each distractors, i.e. each answer accompanying the correct one is aimed at determining the quality of the test question. It is expected of the distractors to be attractive enough for some group of examinees to select them as correct ones. Attention is also drawn to the fact that it is primarily content of questions and instructions that evoke the impression in a candidate's motivation and the manner in which they answer. This, seemingly technical aspect of the education measurement is being analysed by many researchers and tested in conditions of pedagogical experiment in order to extract the most beneficial variations of a given measurement [24].

From among the available results of study it could be concluded that MCQs exams consisting of questions with a lower number of options are characterised by better reliability and validity of measurement. [10, 20, 25]. A good selection procedure should be reliable and that means a high degree of repetitiveness of the results obtained during the exam. A given measuring tool should ensure a result highly independent of features of a person who is taking this measurement. As for the social studies, Earl Babbie pointed to a necessary condition which decides about reliability: impartiality of measurement conditions and precision in scoring [26]. Insufficient reliability of the applied qualification procedure results in low degree of trust if in similar circumstances individual result differ significantly. As can be seen in the presented results, studies of reliability using α-Cronbach's coefficient, exam tests used between the years 2009 – 2013 did not fulfil the assumed criteria of reliability (α coefficient was lower than 0.7). Additionally, the level of reliability of the exams was not dependent on the number of options used in questions that formed the exam set (e.g. Cronbach's α for the exam of 2009 [a four-option] was comparable with the one of 2010 [a five-option]). Inappropriate construction of a tool was of the essence here when considering the low reliability of Nursing Entrance Exams and wrong choice of content in particular, which obviously is just a small part of the whole range of features and properties that are to be evaluated in candidates. Apart from flaws in construction also random errors, i.e. incidental fluctuations of testing results described in a classical theory of a test may lower the value of α coefficient [27]. With a value of $\alpha = 0.5$, random errors are half of variables of the obtained results and measuring in such conditions may only be performed when comparing inter-groups and not while differentiating individually [28]. Low values of the α coefficient incline to review exam tests currently used and based on eliminating elements that differentiate poorly, increasing domain consistency and measuring scale elongation. The last one of the listed elements may be a particularly good method of increasing reliability of the measurement. A greater number of questions in the exam set does not have to – in this case – mean exam time prolongation. As was presented in the meta-analysis results that was published by Aamodt and McShane (1992), in case of a 100-question exam set, students are able to solve on average, around 12.4 problems prepared in three-option MCQs more than they would in a test including four-option MCQs [29]. The authors of the same meta-analysis also state that creating a 100-question test created on the basis of three-option MCQs allow to save even 16 hours of the exam's author's work as opposed to the version based on four-option MCQs [29]. That is why questions with lower number of questions (three instead of four or five options) and less time-consuming may be a good solution in case of the necessity to increase reliability of a test through measuring scale elongation.

A detailed technical analysis of individual test items may become the source of very valuable information concerning the examined group and at the same time, being a complementary part of a

substantive analysis, it remains the basis of evaluation and a foundation on which to build a reliable exam questions database. Previous five-year experience concerning preparing of the MCQ exams allows to determine their weak points: non-equivalence of individual editions, inappropriate selection of questions as far as their simplicity and the differentiating power are concerned. If questions of optimal simplicity are characterised by good differentiating parameters, then they should become the core of the exam questions set and this is the state to be strived at in consecutive years. Creating a good database of exam questions is a long-term process and its resources must be successively renewed due to the phenomenon of exam questions ageing process. This is manifested in the increase in simplicity and decrease of differentiating power revealed in subsequent editions of the exam when it comes to the problems that create a database [3]. A five-year experience in this field allows to elaborate a good test content outline and to prepare a rich database of exam questions from different domains of knowledge and skills, according to strict taxonomy. These activities should ensure a thorough and reliable evaluation of a candidate for nursing studies of the second degree.

Despite the available hard evidence supporting the thesis of better functioning of questions with a lower number of options, there is a strong tendency in the academic circles to use four- and five-option MCQs. One of the reasons of such a state is lack of sufficient knowledge concerning didactic measurement and psychometrics among the academic personnel. This comes as no surprise, given the fact that medical university academics mostly represent professions that do not provide them with pedagogical base. Majority of question authors assumes, rather intuitively, that applying three options should mean a radical increase in the exam simplicity and, as a result, it will overstate the scores of the examinees. However, as Rodriguez concluded in his meta-analysis (2005) [10], the effect of easier guessing in case of a lower number of options is commonly overrated. Numerous results of research proved that lowering a number of options from four / five to three, resulted in the increase of the results within the range of 1.0-1.2% [10, 12, 13, 29]. In the light of the above arguments, it seems that the decision of the organisers of the Nursing Entrance Tests concerning the increase of the number of options in the test questions was not quite rational, which was proved by the statistical analysis of data from the studied period.

5 CONCLUSIONS

Introducing additional option to the test questions did not increase significantly the quality parameters of the Nursing Entrance Exams at MUW. No relevant increase in the selective abilities of the exam or reliability of the measurement were noted. It is recommended to apply four-option questions with simultaneous development of a good test content outline for the needs of future editions of the exam. Also increasing the number of questions in the exam set is worth considering so as to increase reliability and the amount of content covered in the test. Moreover, a team summoned to elaborate the Nursing Entrance Exam at MUW should introduce appropriate guidelines that would improve the selectivity of the exam for the second level of the nursing studies, thus implementing the concept of evidence-based admissions criteria.

REFERENCES

[1] Epstein, R.M. (2007). Medical education - Assessment in medical education. N Engl J Med 356(4), pp. 387-396.

[2] Kulatunga-Moruzi, C. and G.R. Norman. (2002). Validity of admissions measures in predicting performance outcomes: the contribution of cognitive and non-cognitive dimensions. Teach Learn Med 14(1), pp. 34-42.

[3] Case, S.M. and D.B. Swanson. (2002). Constructing Written Test Questions For the Basic and Clinical Sciences. 3rd ed.: National Board of Medical Examiners.

[4] Schuwirth, L.W. and C.P. van der Vleuten. (2003). ABC of learning and teaching in medicine: Written assessment. BMJ 326(7390), pp. 643-5.

[5] McCoubrie, P. (2004). Improving the fairness of multiple-choice questions: a literature review. Med Teach 26(8), pp. 709-12.

[6] Kuncel, N.R. and S.A. Hezlett. (2007). Assessment. Standardized tests predict graduate students' success. Science 315(5815), pp. 1080-1.

[7] Masters, J.C., et al. (2001). Assessment of multiple-choice questions in selected test banks accompanying text books used in nursing education. J Nurs Educ 40(1), pp. 25-32.

[8] Tarrant, M., et al. (2006). The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. Nurse Educ Today 26(8), pp. 662-71.

[9] Vyas, R. and A. Supe. (2008). Multiple choice questions: A literature review on the optimal number of options. Natl Med J India 21(3), pp. 130-133.

[10] Rodriguez, M.C. (2005). Three Options Are Optimal for Multiple-Choice Items: A Meta-Analysis of 80 Years of Research. Educational Measurement: Issues and Practice Summer, pp. 3-13.

[11] Haladyna, T.M. and S.M. Downing. (1993). How many options is enough for a multiple-choice test item? Educational and Psychological Measurement 53(4), pp. 999-1010.

[12] Tarrant, M., J. Ware, and A.M. Mohammed. (2009). An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis. BMC Med Educ 9, pp. 40.

[13] Tarrant, M. and J. Ware. (2010). A comparison of the psychometric properties of three- and four-option multiple-choice questions in nursing assessments. Nurse Educ Today 30(6),pp. 539-43.

[14] Creech, C.J. and C. Aplin-Kalisz. (2011). Developing a selection method for graduate nursing students. J Am Acad Nurse Pract 23(8), pp. 404-409.

[15] Feldt, L.S. (1969). A test of hypothesis that Cronbachs alpha or Kuder-Richardson coefficent 20 is same for 2 tests. Psychometrika 34(3), pp. 363.

[16] Boland, R.J., N.A. Lester, and E. Williams. (2010) Writing Multiple-Choice Questions. Acad Psychiatry 34(4),pp. 310-316.

[17] Downing, S.M. (2005). The effects of violating standard item writing principles on tests and students: the consequences of using flawed test items on achievement examinations in medical education. Adv Health Sci Educ Theory Pract 10(2),pp. 133-43.

[18] Rogers, W.T. and D. Harley. (1999). An empirical comparison of three-and four-choice items and tests: susceptibility to testwiseness and internal consistency reliability. Educ Psychol Meas 59(2),pp. 234-247.

[19] Cizek, G.J. and D.M. O'Day. (1994). Further investigation of nonfunctioning options in multiple-choice test items. Educ Psychol Meas 54(4),pp. 861-872.

[20] Cizek, G.J., K.L. Robinson, and D.M. O'Day. (1998). Nonfunctioning options: A closer look. Educ Psychol Meas 58(4),pp. 605-611.

[21] Haladyna, T.M., S.M. Downing, and M.C. Rodriguez. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. Applied measurement in education 15(3),pp. 309-333.

[22] Downing, S.M. (2003). Validity: on meaningful interpretation of assessment data. Med Educ 37(9),pp. 830-7.

[23] Downing, S.M. and T.M. Haladyna. (2004). Validity threats: overcoming interference with proposed interpretations of assessment data. Med Educ 38(3), pp. 327-33.

[24] Frankfort-Nachmias, C. and D. Nachmias.(2007).Study Guide for Research Methods in the Social Sciences. 7th ed.: Macmillan Higher Education.

[25] Crehan, K.D., T.M. Haladyna, and B.W. Brewer. (1993). Use of an Inclusive Option and the Optimal Number of Options for Multiple-Choice Items. Educ Psychol Meas 53(1), pp. 241-247.

[26] Babbie, E. (2013). The practice of social research. 13th ed.: Belmont: Cengage Learning.

[27] Niemierko, B. (1975). Testy osiągnięć szkolnych. Podstawowe pojęcia i techniki obliczeniowe. 1st ed. Warszawa: Wydawnictwo Szkolne i Pedagogiczne.

[28] Guilford, J.P. (1954). Psychometric methods. 2nd ed. New York: McGraw-Hill.

[29] Aamodt, M. and T. McShane. (1992). A meta-analytic investigation of the effect of various test item characteristics on test scores. Public Pers Manage 21(2),pp. 151-160.